RL hw0

1.(a)

(1). First, prove $V_*(s) \leq max_a Q_*(s,a)$:

$$V_*(s) = V^{\pi_*}(s) = \sum_{a\in\mathcal{A}} \pi_*(a|s) Q^{\pi_*}(s,a)$$

$$\sum_{a\in\mathcal{A}} \pi_*(a|s) Q^{\pi_*}(s,a) \leq max_a Q^{\pi_*}(s,a)$$

$(\because Weighted\ mean\ of\ Q^{\pi}(s,a)\ over\ a\ \leq\ the\ maximum)$

$max_a Q^{\pi_*}(s,a) \;=\; max_a Q_*(s,a)$ ■

Then, show $V_*(s) < max_a Q_*(s,a)$ cannot happen by contradiction.

Let's define a policy $\pi'(a|s) = \begin{cases} 1,\ if\ a = argmax_{a'} Q_*(s,a') \\ 0,\ else \end{cases}$.

Suppose $V_*(s) < max_a Q_*(s,a)$, then

$$V^{\pi'}(s) = \sum_{a\in\mathcal{A}} \pi'(a|s) Q_*(s,a) = max_a Q_*(s,a) > V_*(s)$$

But $V_*(s) = max_\pi V^{\pi}(s)$, contradiction with $V^{\pi'}(s) > V_*(s)$.

Thus, $V_*(s) = max_a Q_*(s,a)$ ■

(2). By leveraging the fact that $Q^{\pi}(s,a) = R_s^a + \gamma \sum_{s'} P_{ss'}^a V^{\pi}(s')$.

We can show:

$$Q_*(s,a) = Q^{\pi_*}(s,a) = R_s^a + \gamma \sum_{s'} P_{ss'}^a V^{\pi_*}(s')$$

$$= R_s^a + \gamma \sum_{s'} P_{ss'}^a V^*(s') \;■$$

1.(b)

For any two action-value functions Q, Q':

$\|T^*(Q) - T^*(Q')\|_\infty = max_{(s,a)} |[T^*(Q)](s,a) - [T^*(Q')](s,a)|$

$$= max_{(s,a)} \left| \left[R_s^a + \gamma \sum_{s\prime} P_{ss\prime}^a max_{a\prime} Q(s', a') \right] - \left[R_s^a + \gamma \sum_{s\prime} P_{ss\prime}^a max_{a\prime} Q'(s', a') \right] \right|$$

$$= max_{(s,a)} \left| \gamma \sum_{s\prime} P_{ss\prime}^a max_{a\prime} Q(s', a') - \gamma \sum_{s\prime} P_{ss\prime}^a max_{a\prime} Q'(s', a') \right|$$

$$\leq max_{(s,a)} \gamma | \; max_{s\prime} max_{a\prime} Q(s', a') - max_{s\prime} max_{a\prime} Q'(s', a') |$$

$$\leq max_{(s,a)} max_{s\prime} max_{a\prime} \gamma | \; Q(s', a') - Q'(s', a') |$$

$$= max_{(s\prime,a\prime)} \gamma | \; Q(s', a') - Q'(s', a') |$$

$$= max_{(s,a)} \gamma | \; Q(s, a) - Q'(s, a) |$$

$$= \gamma \| Q - Q' \|_\infty \blacksquare$$

Proved that $\| T^*(Q) - T^*(Q') \|_\infty \leq \gamma \| Q - Q' \|_\infty$, so T^* is a γ -contraction operator in terms of ∞ −norm.

2.

$$L(\pi) = \sum_{a \in \mathcal{A}} \left(\pi(a|s) Q_\Omega^{\pi_k}(s, a) - \pi(a|s) \log \pi(a|s) \right) - \mu \left(\sum_{a \in \mathcal{A}} \pi(a|s) - 1 \right)$$

For each $a \in \mathcal{A}$,

$$\frac{\partial L(\pi)}{\partial \pi(a|s)} = Q_\Omega^{\pi_k}(s, a) - \log \pi(a|s) - 1 - \mu = 0$$

$$\Rightarrow \log \pi(a|s) = Q_\Omega^{\pi_k}(s, a) - 1 - \mu$$

$$\Rightarrow \pi(a|s) = e^{Q_\Omega^{\pi_k}(s,a) - 1 - \mu}$$

Because of the constraint $\sum_{a \in \mathcal{A}} \pi(a|s) - 1 = 0$

$$\sum_{a \in \mathcal{A}} \pi(a|s) = e^{-1-\mu} \sum_{a \in \mathcal{A}} e^{Q_\Omega^{\pi_k}(s,a)} = 1$$

$$\Rightarrow e^{-1-\mu} = \frac{1}{\sum_{a \in \mathcal{A}} e^{Q_\Omega^{\pi_k}(s,a)}}$$

$$\Rightarrow e^{1+\mu} = \sum_{a \in \mathcal{A}} e^{Q_{\Omega}^{\pi_k}(s,a)}$$

$$\Rightarrow \mu = \ln \sum_{a \in \mathcal{A}} e^{Q_{\Omega}^{\pi_k}(s,a)} - 1$$

Then plug into $\pi(a|s) = e^{Q_{\Omega}^{\pi_k}(s,a)-1-\mu}$

$\pi(a|s) = e^{Q_{\Omega}^{\pi_k}(s,a)-1-\mu} = \frac{exp\left(Q_{\Omega}^{\pi_k}(s,a)\right)}{exp\left(\sum_{a \in \mathcal{A}} e^{Q_{\Omega}^{\pi_k}(s,a)}\right)}$ is the optimal solution.

Thus,

$$\pi_{k+1}(\cdot\,|s) = argmax_{\pi}\left\{\langle \pi(\cdot\,|s), Q_{\Omega}^{\pi_k}(s,\cdot) \rangle - \Omega\big(\pi(\cdot\,|s)\big)\right\}$$

$$= \frac{exp\left(Q_{\Omega}^{\pi_k}(s,a)\right)}{exp\left(\sum_{a \in \mathcal{A}} e^{Q_{\Omega}^{\pi_k}(s,a)}\right)} \blacksquare$$